JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest

Fachrul Mustofa [1], Achmad Nuruddin Safriandono [2], Ahmad Rofiqul Muslikh [3] and De Rosal Ignatius Moses Setiadi [1,*]

[1] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Central Java 50131, Indonesia; e-mail : 111201710555@mhs.dinus.ac.id, moses@dsn.dinus.ac.id
[2] Faculty of Engineering, Sultan Fatah University, Demak, Central Java 59516, Indonesia; e-mail : udinozz@gmail.com
[3] Faculty of Information Technology, University of Merdeka, Malang, East Java 65146, Indonesia; e-mail : rofickachmad@unmer.ac.id
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: Diabetes Mellitus is a hazardous disease, and according to the World Health Organization (WHO), diabetes will be one of the main causes of death by 2030. One of the most popular diabetes datasets is PIMA Indians, and this dataset has been widely tested on various machine learning (ML) methods, even deep learning (DL). But on average, ML methods are not able to produce good accuracy. The quality of the dataset and features is the most influential thing in this case, so deeper investment is needed to examine this dataset. This research will analyze and compare the PIMA Indians and Abelvikas datasets using the Random Forest (RF) method. The two datasets are imbalanced, in fact, the Abelvikas dataset is more imbalanced and has a larger number of classes so it is be more complex. The RF was chosen because it is one of the ML methods that has the best results on various diabetes datasets. Based on the test results, very contrasting results were obtained on the two datasets. Abelvikas had accuracy, precision, and recall, reaching 100%, and PIMA Indians only achieved 75% for accuracy, 87% for precision, and 80% for the best recall. Testing was done with 3, 5, 7, 10, and 15 tree number parameters. Apart from that, it was also tested with k-fold validation to get valid results. This determines that the features in the Abelvikas dataset are much better because more complete glucose features support them.

Keywords: Classification Diabetes Types; Comprehensive analysis for diabetes types classification; Prediction for health technology; Random Forest; Feature Analysis; Abelvikas Dataset.

Received: 15th July 2023
Revised: 6th August 2023
Accepted: 7th August 2023
Published: 8th August 2023

1. Introduction

Diabetes Mellitus is a chronic disease in which the sufferer's body has high blood sugar levels. Diabetes occurs because the pancreas cannot make the insulin hormone properly so that it cannot be used in the body efficiently[1]. Diabetes is a dangerous disease. From predictions made by the World Health Organization (WHO), diabetes is one of the leading causes of death in 2030, and the death rate will increase by 54% between 2015 and 2030 in America despite medical advances and prevention efforts[2]. There is a need for early detection of this disease. If not treated quickly and appropriately, it can cause many complications. Short-term complications include ketoacidosis, coma, and death. Meanwhile, long-term complications that can arise include a malfunctioning heart that causes stroke, damage to the retina in the eye, chronic kidney failure, nerves, and teeth [3].

There are two types of diabetes, namely type 1 and type 2 diabetes. Although the symptoms are similar, there are differences such as the causes and treatment methods. For type 1 diabetes, the hormone insulin cannot be produced by the body because the pancreas's beta cells are damaged, resulting in decreased insulin production. Whereas for type 2 diabetes, the body usually produces insulin, the cells in the body are less sensitive, so it cannot be used optimally[4], [5]. In the traditional paradigm, type 2 diabetes is only experienced by adults,

and type 1 is only experienced in children. Still, it is inaccurate because all age groups can experience this disease. The appearance of type 1 diabetes symptoms in adults may differ from those in children. However, difficulties in diagnosis can occur in children, adolescents, and adults[5]. Technological developments in various sectors are increasing[6], [7], especially in the health sector. Data mining technology has been applied in this field to help solve problems. Data mining has several techniques, including classification, clustering, association, and regression. Classification is a technique in data mining that is used to classify data based on the relationship of the data to sample data. Classification techniques in data mining can be used to help identify types of diabetes based on existing data.

Diabetes mellitus classification can be done using machine learning methods, for example, classification and regression tree (CART), random forest (RF), logistic regression (LR), decision tree (DT), and support vector machine (SVM). The advantages of the CART algorithm include more accurate results, faster calculations, easier to interpret, and can be used for large data sets. The downside is that CART is less stable and sensitive to new data. So it really depends on the number of samples, if the sample data changes then the decision tree results will change[8]. The advantages of RF include handling missing values (missing data) in the dataset, producing lower errors, effectively handling large amounts of training data efficiently, providing good classification results, and avoiding overfitting [9]. Previous research [10], which compared the CART and RF algorithms, found that the accuracy of the CART method was 64.9%, while the RF was 71%. This shows that the RF method is superior. Another study [11] also compared LR, DT, SVM, and RF for diabetes classification. In general, LR is a simple probability-based classification method. The LR is superior for cases when the relationship between features and targets is linear or almost linear, but is ineffective when the relationship between features and targets is not linear is susceptible to overfitting if there is no regulation and cannot handle different features, highly non-linear or complex feature interactions. SVM is effective over high-dimensional feature spaces, suitable for cases with little training data, and can handle non-linear classification problems through kernels. However, SVM is weak if the dataset is large and requires selecting appropriate kernels. The DT is superior because it does not require normalization or scaling of the data and is suitable for data that has categorical features, but if the tree is too deep, it is susceptible to overfitting and can be unstable if data changes are small[12]. The results of the research [11] produced the highest accuracy in DT and RF.

Another study [13] also tested several popular ML methods such as DT, RF, SVM, and K-nearest neighbor (KNN) on private datasets taken from the Department of Medical Services, Bangkok, between 2019 and 2021. The result was that RF was also the best method in two types of experiments, where the first experiment was classification without construct interaction terms and with interaction terms. What is meant by interaction terms here is optimization with feature selection and hyperparameter tuning. The RF produces 88.2% accuracy without interaction terms, while accuracy increases to 97.5% with interaction terms. Diabetes classification research [14] also compared RF and DT methods. This research explains that the dataset consists of 19 variables for 403 out of the 1,046 surveyed topics among African Americans in a study aimed at determining the prevalence of obesity, diabetes, and other cardiovascular risk factors in Central Virginia. The result is that the RF method is also superior to DT, with an accuracy of 86.53% without feature selection and an accuracy of 92.02% with feature selection.

Features influence classification results in data mining [7], so feature selection is very important[15], [16]. This can be done if the number of features is large enough and complete. Popular datasets such as PIMA Indians have minimal features, namely eight features and one class attribute[17]. , in this case perhaps feature selection is not effective enough. However, if a dataset is designed with knowledge of the most influential features, it will improve it and produce the best performance with various algorithms. However, in data mining, the dataset as input is the thing that most determines the next handling steps. Various studies such as[18]–[21], used the PIMA Indians dataset on various ML classifiers. In research [18], several features have been tested, namely 3, 5, and all features. The result is that each number of features can be superior in one of the sections, namely recall, specificity, F1 score, and area under the curve (AUC). Furthermore, based on test results from research [18]–[21], the ML method tested could not achieve an accuracy of up to 90%. This contrasts sharply with the Abelvikas dataset tested in research [11], where classification accuracy reached 100%. This 100% result sometimes makes other researchers less confident and may need to retest.

Based on the literature above, this research aims to further analyze the performance of the RF method on two datasets, namely PIMA Indians and Abelvikas. These two datasets have a number of records that are not much different but have very different features. The RF method was chosen because it is able to produce the best accuracy on various datasets. The contributions of this paper are:

1. Further analyze the RF method by setting the parameters for the number of k-fold cross-validations and limiting the number of trees.
2. Analyze what features and their influence on classification performance on both datasets.
3. Analyze in more detail based on accuracy, precision, and recall measuring instruments, which measuring instrument is most suitable to be used as a benchmark, the most important reasons.

The rest of this paper is presented in three sections, namely methodology in the second section discusses the theory, literature, methods used, and the reasons. The third section explains the results and analysis; the last is the conclusion.

2. Methodology

This section begins by explaining the research methodology presented in Figure 1. The methodology used in this research is quite simple, consisting of four main stages: dataset collection, preprocessing, RF classification, and evaluation, presented in subsections 2.1 to 2.4.

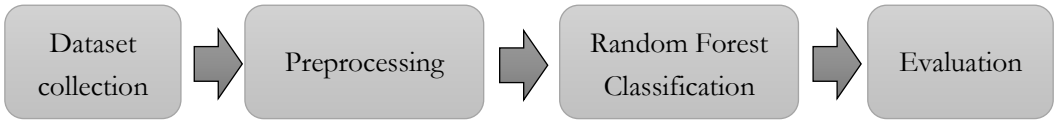

Figure 1. Research Methodology.

2.1 Dataset Collection

In this study, two datasets were used, namely the Abelvikas dataset and the PIMA Indians dataset. The Abelvikas dataset was chosen because in research [11], the RF method was able to produce accuracy up to 100%, so we reviewed it because in most diabetes datasets, especially in the PIMA Indians dataset, it is very difficult to produce accuracy up to 90% if done using standard ML methods. Based on Abelvikas data downloaded from the link data.world/abelvikas/diabetes-type-dataset. It is known that there are 1009 records with eight attributes, including class and type. The attributes of the Abelvikas dataset are presented in Table 1. The PIMA Indians dataset can be accessed at the url: www.kaggle.com/uciml/pima-indians-diabetes-database, this dataset has 768 records (500 non-diabetics and 268 diabetics). The features used in the PIMA Indians dataset are presented in Table 2.

Table 1. Abelvikas dataset attributes.

No	Attribute	Data Type	Note
1	Age	Numeric	Patient age
2	BS Fast	Numeric	Fasting blood sugar before eating (mmol/L)
3	BS pp	Numeric	Blood sugar within 90 minutes after eating (mmol/L)
4	Plasma R	Numeric	Randomized plasma glucose test at any time (mmol/L)
5	Plasma F	Numeric	A fasting plasma glucose test is usually done in the morning (mmol/L)
6	HbA1c	Numeric	This test is performed to measure the average blood glucose level over the last 2-3 months (mmol/L)
7	Type	String	Diagnosis type (Type 1, Type 2, and Normal)
8	Class	Boolean	Diagnostic results (true or false)

Table 2. PIMA Indians dataset attributes.

No	Attribute	Data Type	Note
1	Preg	Numeric	The number of pregnancies
2	Gluc	Numeric	Glucose plasma levels two hours after consuming glucose
3	BP	Numeric	Diastolic blood pressure (mm Hg)
4	Skin	Numeric	Thickness of the skin fold on the triceps of the upper arm (mm)
5	Insulin	Numeric	Insulin serum levels in the blood two hours after the glucose test (lh/ml)
6	BMI	Numeric	Body mass index [weight in kg/(Height in m)], an index used to evaluate a person's relative weight
7	DPF	Numeric	Diabetes pedigree function is a value that measures genetic risk factors based on a family history of diabetes.
8	Age	Numeric	patient's age in years
9	Class	Boolean	Diagnostic results (true or false)

Based on the data presented above, it appears that there are quite a lot of significant differences between these two datasets. The Abelvikas dataset has six main features that have numeric values, while the other two are label attributes. One attribute, namely age, is shared by both datasets. There are five attributes directly related to the patient's blood for Abelvikas, whereas there are four for PIMA Indians. The PIMA Indians dataset only has two labels, namely diabetic and non-diabetic, while Abelvikas has three labels, namely diabetic type 1, diabetic type 2, and non-diabetic/ normal.

2.2 Preprocessing

In this research, several preprocessing was carried out, namely first, by eliminating duplicate data. Duplicate data is removed so that the data is cleaner, and no data is the same, so the data mining results will be more accurate and faster. At this stage, we do it manually using MS Excel. In the Abelvikas dataset, 386 duplicate data were obtained, so only 623 records were left, see Figure. 1. Meanwhile, in the PIMA Indians dataset, no duplicate data was found.

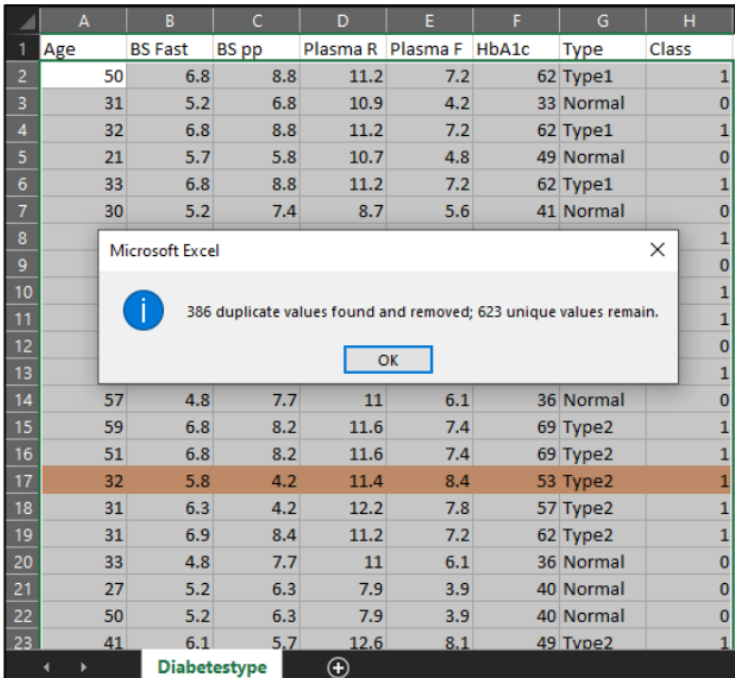

Figure 1. Search and delete duplicate data.

Then, the next preprocessing step is to remove unused attributes. The Abelvikas dataset has two string attributes, so the class attribute was removed. The type attribute is the result of the diagnosis type (normal, type 1, or type 2), and this attribute is retained because it has a more specific diagnosis. Next, each record was identified based on type, of which there were 428 normal types, 101 Type 1 data, and 94 Type 2 data. From this analysis, it was concluded that the Abelvikas and PIMA Indians datasets were imbalance datasets. Even more specifically, the Abelvikas dataset was more imbalanced, more clearly see Figure 2. According to [22], [23], classification problems on imbalanced datasets have several significant implications. One is accuracy imbalance, where models built on such datasets tend to have high accuracy overall but can be very ineffective at predicting minority classes. This happens because the model tends to predict the majority class for almost all data, while minority classes are often ignored. Additionally, this can result in the model being biased towards the majority class, ignoring important patterns in the minority class. Model evaluation is also problematic, as metrics such as accuracy no longer provide an accurate picture of model performance. Finally, handling data rarely appearing in minority classes is difficult due to their small number, which can result in the model making more errors in predicting those classes.

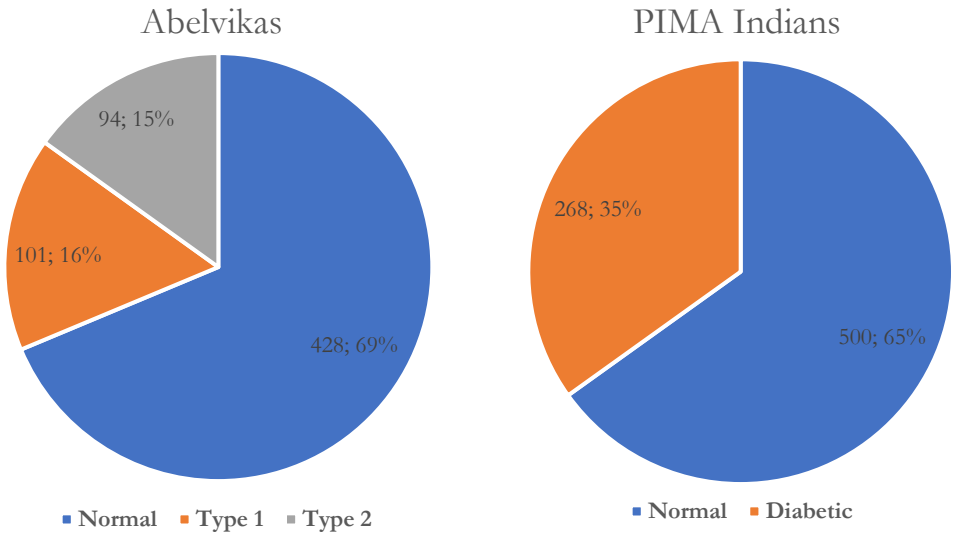

Figure 2. Dataset composition based on class labels.

2.3 Classification Method

The RF is the method proposed in this research, and this algorithm is relatively good to handling imbalanced dataset problems. RF is ensemble learning consisting of many decision trees. Each tree is learned from samples taken with replacements from the dataset. When RF makes predictions, each tree votes and the final result is the result of the majority of all trees. RF randomly selects a subset of attributes for each tree, which helps prevent the model from being overly influenced by the majority of attributes. This can reduce bias towards the majority. RF has also been tested on various diabetes datasets, including the two datasets above, and the results are quite good. But there has been no in-depth discussion regarding comparing these two datasets. In this study, the diabetes dataset was separated into training and testing data. The k-fold cross-validation split method is used to obtain more valid results, where k values of 3, 5, 7 and 10 fold are used.

The RF has more than one tree, and each tree will be generated based on the training process. The number of trees is one of the important parameters in RF, and is included in the category of hyperparameters that can be adjusted. The number of trees will affect RF performance. Increasing the number of trees in an RF can improve model performance. With a larger number of trees, RF can better generalize and reduce overfitting on training data and have more stable predictions. However, if the number of RF trees is too large or unlimited, RF continues to create new trees until it meets certain stopping conditions. This can cause RF to be very complex and slow in training. Without limitation, RF also has the potential to overfit the training data. So, in this study, the number of trees is limited. It is limited to 3, 5, 7, 10 and 15 for each k-fold.

2.4 Evaluation

In this research, the evaluation of classification methods was carried out using several measuring tools such as accuracy, precision, and recall. Accuracy is used to measure the extent to which the classification model is correct in predicting all classes correctly, which can be calculated by Eq. (1). precision is used to measure the extent to which the model's positive predictions are correct from all the positive predictions of the classification results, this can be calculated by Eq. (2). Meanwhile, recall is useful for measuring the extent to which the method is able to identify all true positive instances from all existing positive instances, which is calculated by Eq. (3).

$$acc = \frac{All\ true\ prediction}{All\ Data} \tag{1}$$

$$pre = \frac{True\ positive}{All\ Positive\ data} \tag{2}$$

$$rec = \frac{True\ positive}{True\ positive + False\ negative} \tag{3}$$

Where all true data is true positive and true negative, all positive data is true positive and false positive, and all data contains true positive, true negative, false positive, and false negative.

3. Results and Analysis

This research was implemented using the Python programming language. The scikit learn library uses the RandomForestClassifier class and the Gini index to form a tree. The number of trees is set with the n_estimators parameter. Meanwhile, to divide data by k-fold using the StratifiedKFold function, where the n_splits parameter is used to determine the number k = 3,5, or 10. StratifiedKFold was chosen because the dataset used is imbalanced, and this function is relatively better for imbalanced data distribution. The results of the method implementation are presented in Table 3 for the Abelvikas dataset and Table 4 for the PIMA Indians dataset.

Table 3. Results in Abelvikas Dataset.

K-Fold	Num of Trees	Accuracy	Precision	Recall
3-Fold Cross Validation	3	99.357	98.679	98.639
	5	99.518	99.262	98.969
	7	99.839	99.669	99.649
	10	100	100	100
	15	100	100	100
5-Fold Cross Validation	3	99.518	98.96	99.268
	5	99.839	99.922	99.649
	7	100	100	100
	10	100	100	100
	15	100	100	100
10-Fold Cross Validation	3	99.839	99.669	99.649
	5	99.839	99.922	99.649
	7	100	100	100
	10	100	100	100
	15	100	100	100

Table 4. Results in PIMA Indians Dataset.

K-Fold	Num of Trees	Accuracy	Precision	Recall
3-Fold Cross Validation	3	72.265	79.000	78.528
	5	72.005	81.200	77.039
	7	72.265	82.800	76.525
	10	74.218	86.400	76.868
	15	76.041	83.800	80.268
5-Fold Cross Validation	3	69.921	78.200	76.218
	5	70.963	79.000	76.998
	7	73.046	81.600	78.011
	10	75.651	87.000	78.096
	15	75.781	84.200	79.734
10-Fold Cross Validation	3	72.265	81.800	77.024
	5	75.390	83.400	79.732
	7	76.041	85.800	79.151
	10	75.651	88.200	77.504
	15	75.520	85.800	78.571

Based on the data produced in the two tables, both datasets have striking results based on accuracy, precision, and recall. The test results on the Abelvikas dataset look very good, with accuracy, precision, and recall all above 99% or even up to 100%. Meanwhile, in the PIMA Indians dataset, the accuracy, precision, and recall vary, namely around 69 to 75% for accuracy, 78% to 87% for precision, and 76% to 80% for recall. These results are identical to previous research, either with the same or a different dataset. This shows that the RF performance for classifying the PIMA Indians dataset is proven to be valid, as well as the Abelvikas dataset.

The unequal results presented in Tables 3 and 4 are of course greatly influenced by the features in the dataset. Because both datasets are imbalanced and the number of records is slightly different. Even the Abelvikas dataset tends to be more imbalanced and has more classes. Surprisingly, the classification results show that RF is more powerful for predicting the Abelvikas dataset. This is caused by the dominance of the blood glucose feature in the Abelvikas dataset, i.e., five of the six features. These features may be particularly informative in differentiating the three types of diabetes diagnosis. Blood glucose test results and patient age are important factors in the diagnosis of diabetes.

A more detailed discussion needs to be discussed on these three measuring instruments. It should be noted that these three measuring instruments are the most popular and have been widely used in previous research. Accuracy is used to measure the extent to which the model can correctly predict both the positive class (patients who have the disease) and the negative class (patients who do not have the disease). Accuracy is an important metric in general, but in cases of class imbalance (where the number of patients with diabetes may be fewer), accuracy may not provide an accurate picture of model performance. Accuracy will tend to predict the majority class, so even though accuracy is high, it will not necessarily be able to detect diabetes patients. Meanwhile, precision measures the extent to which positive predictions from the model are correct. In the context of disease prediction such as diabetes, precision measures what percentage of patients predicted to have the disease actually have the disease. Precision is essential when false positives can have serious consequences or when the cost of further testing is high.

Recall is also known as sensitivity or true positive rate, useful for measuring the extent to which a model can detect all true positive cases. In the context of disease prediction, recall measures how well the model can detect all patients who actually have the disease. Recall is necessary when one does not want to miss true disease cases, even if it means there are some false positives [24]. Thus, in the context of disease prediction such as diabetes, recall is often considered a more important metric than accuracy and precision. This is because identifying

all patients who actually have the disease is a top priority, and we want to avoid false negative errors that can have serious consequences.

4. Conclusions

This research has succeeded in carrying out the objectives of this research. The RF method obtained satisfactory and appropriate prediction results, especially on the Abelvikas dataset. The results regarding k-fold validation and the number of trees show that k-fold validation has a smaller message than the number of trees in this case. Using more trees produces better performance based on accuracy, precision, and recall. The Abelvikas dataset was also proven to have better features and minimum data noise, even though it initially had many duplicate data. However, after deleting the performance data on the Abelvikas dataset, it is much better than PIMA Indians. The Abelvikas dataset is relatively more imbalanced with a better number of classes, but the classification results are very unequal with significant advantages. Of course, this is influenced by the quality and features used in the Abelvikas dataset. The number of blood glucose features is the majority in the Abelvikas dataset, this allows for better results. Finally, in the case of health prediction, you need to know that recall is a more important measuring tool, especially when it can occur in imbalanced datasets. In the future, diabetes dataset collection will be better if it has richer glucose features to produce the best accuracy and a relatively simple method based on ML.

Author Contributions: Conceptualization: F. Mustofa and D.R.I.M. Setiadi; methodology: all authors; software: F. Mustofa; validation: all authors; formal analysis: F. Mustofa, A. N. Safriandono and A. R. Muslikh; investigation: F. Mustofa and D.R.I.M. Setiadi; resources: F. Mustofa; data curation: F. Mustofa and A. N. Safriandono; writing—original draft preparation: F. Mustofa; writing—review and editing: D.R.I.M. Setiadi and A. R. Muslikh; visualization: F. Mustofa and A. N. Safriandono; supervision: D.R.I.M, Setiadi, A. N. Safriandono, and A. R. Muslikh; project administration: A. N. Safriandono, and A. R. Muslikh.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*, Elsevier Inc., 2020, pp. 327–339. doi: 10.1016/B978-0-12-819061-6.00014-8.

[2] W. R. Rowley, C. Bezold, Y. Arikan, E. Byrne, and S. Krohe, "Diabetes 2030: Insights from Yesterday, Today, and Future Trends," *Popul. Health Manag.*, vol. 20, no. 1, pp. 6–12, Feb. 2017, doi: 10.1089/pop.2015.0181.

[3] H. Das, B. Naik, and H. S. Behera, "Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach," in *Advances in Intelligent Systems and Computing*, vol. 710, Springer Verlag, 2018, pp. 539–549. doi: 10.1007/978-981-10-7871-2_52.

[4] E. Pekel Özmen and T. Özcan, "Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm," *J. Forecast.*, vol. 39, no. 4, pp. 661–670, Jul. 2020, doi: 10.1002/for.2652.

[5] American Diabetes Association, "Classification and Diagnosis of Diabetes," *Diabetes Care*, vol. 38, no. Supplement_1, pp. S8–S16, Jan. 2015, doi: 10.2337/dc15-S005.

[6] B. M. P. Waseso and N. A. Setiyanto, "Web Phishing Classification using Combined Machine Learning Methods," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 11–18, Aug. 2023, doi: 10.33633/jcta.v1i1.8898.

[7] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, "Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023, doi: 10.33633/jcta.v1i1.9185.

[8] R. J. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis," in *Annual Meeting of the Society for Academic Emergency Medicine*, 2000, no. 310, p. 14p. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf

[9] B. Boehmke and B. Greenwell, "Random Forests," in *Hands-On Machine Learning with R*, Chapman and Hall/CRC, 2019, pp. 203–219. doi: 10.1201/9780367816377-11.

[10] H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, "A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes," *J. Res. Health Sci.*, vol. 18, no. 2, 2018.

[11] O. Adigun, F. Okikiola, N. Yekini, and R. Babatunde, "Classification of Diabetes Types using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 152–161, 2022, doi: 10.14569/IJACSA.2022.0130918.

[12] B. Boehmke and B. Greenwell, *Hands-On Machine Learning with R*. Chapman and Hall/CRC, 2019. doi: 10.1201/9780367816377.

[13] M. Phongying and S. Hiriote, "Diabetes Classification Using Machine Learning Techniques," *Computation*, vol. 11, no. 5, p. 96, May 2023, doi: 10.3390/computation11050096.

[14] K. K. Chari, M. C. Babu, and S. Kodati, "Classification of Diabetes using Random Forest with Feature Selection Algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 1295–1300, Nov. 2019, doi: 10.35940/ijitee.L3595.119119.
[15] E. H. Rachmawanto, D. R. Ignatius Moses Setiadi, N. Rijati, A. Susanto, I. U. Wahyu Mulyono, and H. Rahmalan, "Attribute Selection Analysis for the Random Forest Classification in Unbalanced Diabetes Dataset," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2021, pp. 82–86. doi: 10.1109/iSemantic52711.2021.9573181.
[16] D. R. Ignatius Moses Setiadi *et al.*, "Effect of Feature Selection on The Accuracy of Music Genre Classification using SVM Classifier," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2020, pp. 7–11. doi: 10.1109/iSemantic50169.2020.9234222.
[17] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clin. eHealth*, vol. 4, pp. 12–23, Jan. 2021, doi: 10.1016/j.ceh.2020.11.001.
[18] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.
[19] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.
[20] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.
[21] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," *Diagnostics*, vol. 13, no. 14, p. 2383, Jul. 2023, doi: 10.3390/diagnostics13142383.
[22] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013. [Online]. Available: https://ieeexplore.ieee.org/book/6542371
[23] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study1," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Nov. 2002, doi: 10.3233/IDA-2002-6504.
[24] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, Aug. 2001, doi: 10.1016/S0933-3657(01)00077-X.